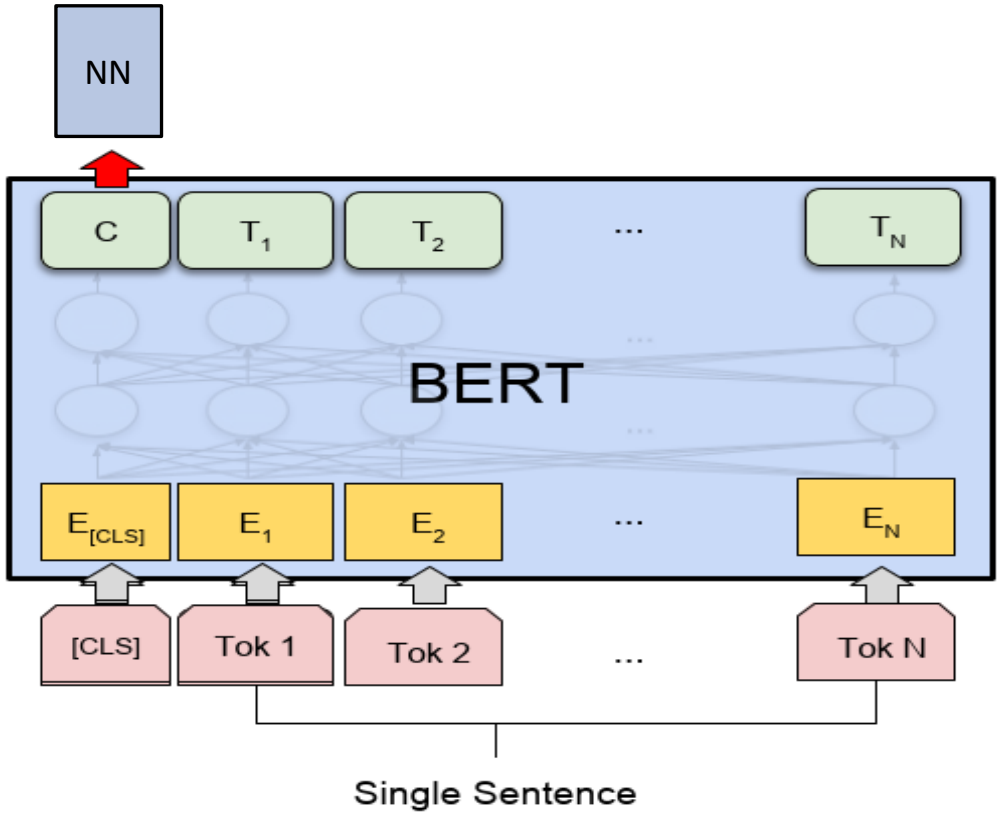
Outline

- Supervised Learning
  - BERT
- Semi-supervised Learning
  - Meta Pseudo Labels
- Results

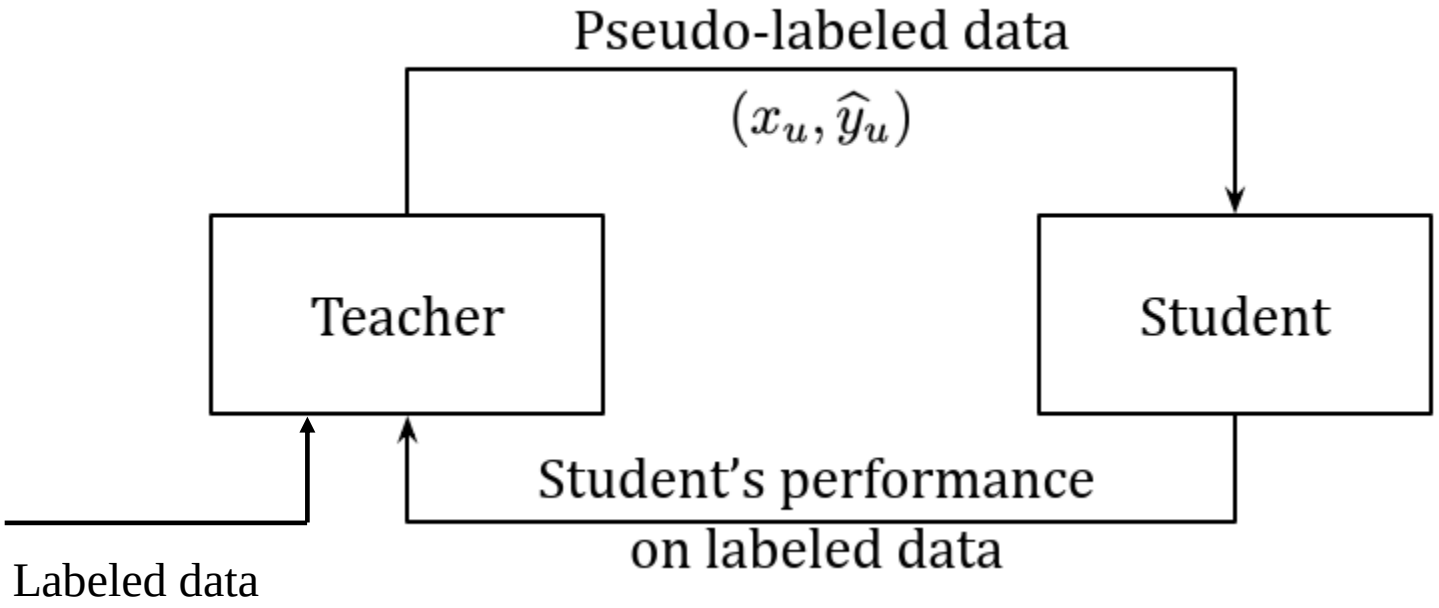
Movie Review

- Total 10662 samples
- 5331 positive samples
  - e.g. worth the effort to watch .
- 5331 negative samples
  - e.g. simplistic , silly and tedious .

Supervised Learning



Meta Pseudo Labels



Semi-supervised Learning

- Meta Pseudo Labels

θ_T: Teacher network parameters

θ_S: Student network parameters

For student

$$\theta_S^{PL} = argmin_{\theta_S} \ \mathcal{L}_u(\theta_T, \theta_S) \ ,$$

$$\mathcal{L}_u(\theta_T, \theta_S) = \ \mathbb{E}_{x_u}\left[\mathrm{CE}\big(T(x_u;\theta_T), S(x_u;\theta_S)\big)\right]$$

For teacher

$$min_{\theta_T} \ \mathcal{L}_l(\theta_S^{PL}(\theta_T)) = min_{\theta_T} \ \mathbb{E}_{x_l, y_l}\left[\mathrm{CE}\big(y_l, S(x_l;\theta_S^{\mathrm{PL}})\big)\right]$$

Meta Pseudo Labels

For teacher

$$min_{\theta_T} \ \mathcal{L}_l(\theta_S^{PL}(\theta_T)) = min_{\theta_T} \mathbb{E}_{x_l, y_l}\left[\mathrm{CE}\left(y_l, S(x_l; \theta_S^{\mathrm{PL}})\right)\right]$$

$$\theta_S^{\mathrm{PL}}(\theta_T) \approx \theta_S - \eta_S \cdot \nabla_{\theta_S} \mathcal{L}_u\left(\theta_T, \theta_S\right)$$

$$\min_{\theta_T} \quad \mathcal{L}_l\left(\theta_S - \eta_S \cdot \nabla_{\theta_S} \mathcal{L}_u\left(\theta_T, \theta_S\right)\right)$$

Meta Pseudo Labels

Update student

$$\theta'_S = \theta_S - \eta_S \nabla_{\theta_S} \mathcal{L}_u(\theta_T, \theta_S)$$

Update teacher

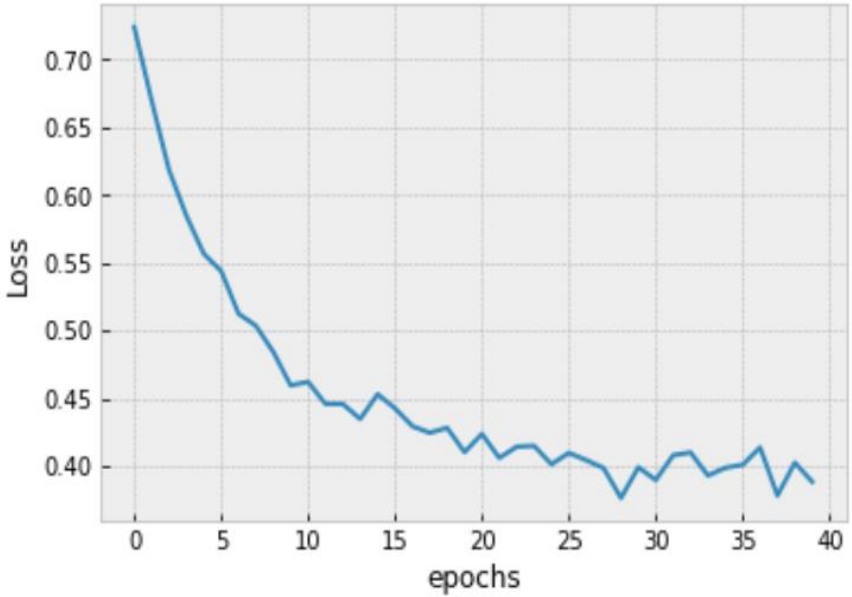
$$\theta'_T = \theta_T - \eta_T \nabla_{\theta_T} \mathcal{L}_l\big(\ \theta_S - \nabla_{\theta_S} \mathcal{L}_u(\theta_T, \theta_S)\ \big)$$

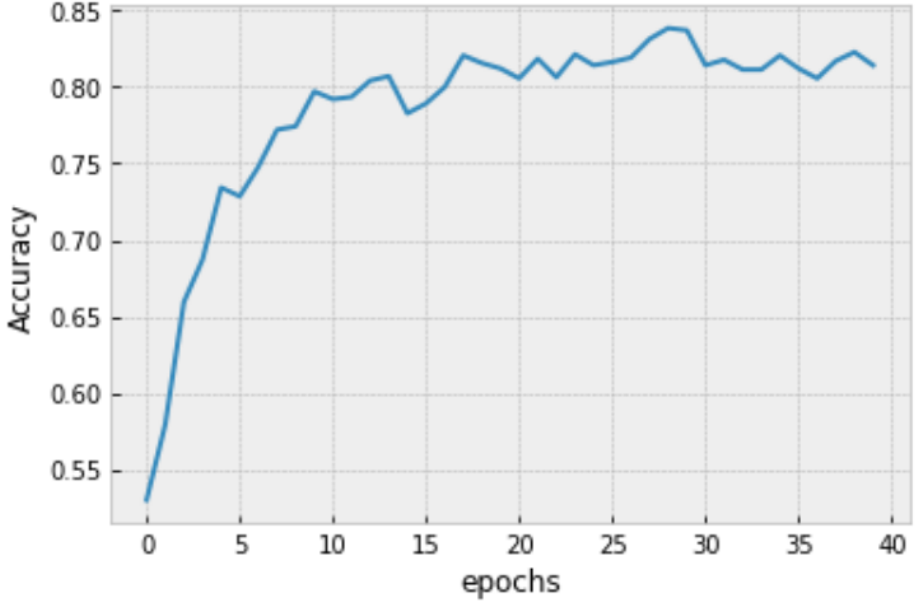
Movie Review

- Total 10662 samples
- 1024 labeled data for training
- 256 labeled data for test
- Others used for unlabeled data

Results

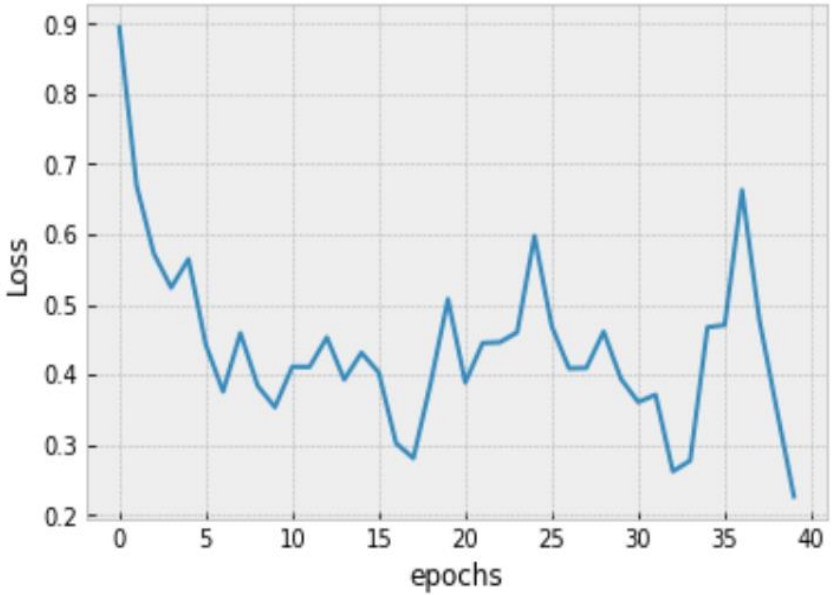
Supervised

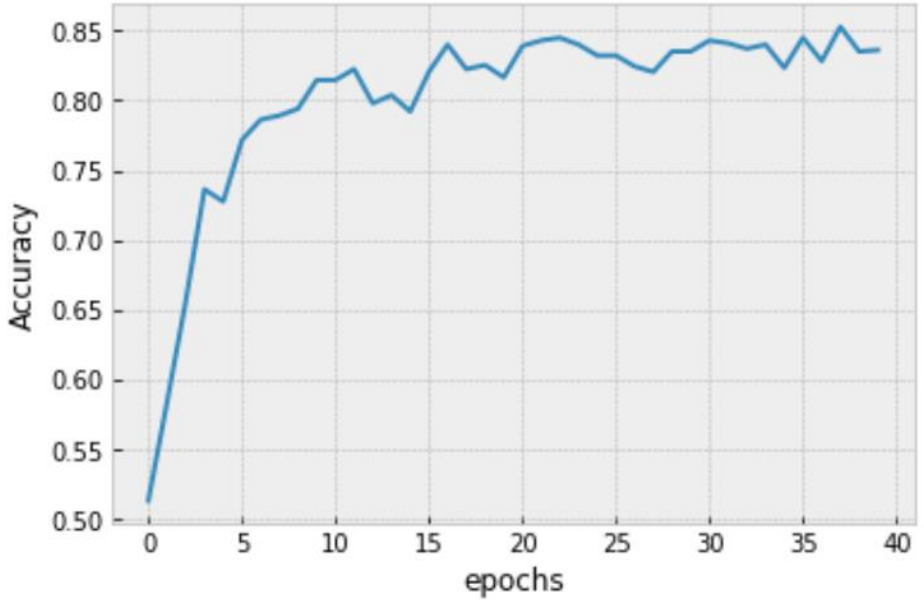




Results

MPL – student on labeled data





Results

On test set

Supervised

- accuracy 0.8518 ± 0.0472

Semi-supervised

- accuracy 0.8477 ± 0.0513